УПРАВЛЕНИЕ В СОЦИАЛЬНЫХ И ЭКОНОМИЧЕСКИХ СИСТЕМАХ

УДК 004.75

РАЗРАБОТКА МЕТОДА ПРОАКТИВНОГО ОБНАРУЖЕНИЯ МОШЕННИЧЕСТВА ПОТРЕБИТЕЛЕЙ УСЛУГ ТЕЛЕКОММУНИКАЦИОННОЙ КОМПАНИИ[1]

Статья поступила в редакцию 14.11.2016, в окончательном варианте – 25.11.2016.

Нгуен Туан Ань, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 28, e-mail: anhtuank37@gmail.com

Щербаков Максим Владимирович, доктор технических наук, главный научный сотрудник, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 28, e-mail: maxim.shcherbakov@vstu.ru

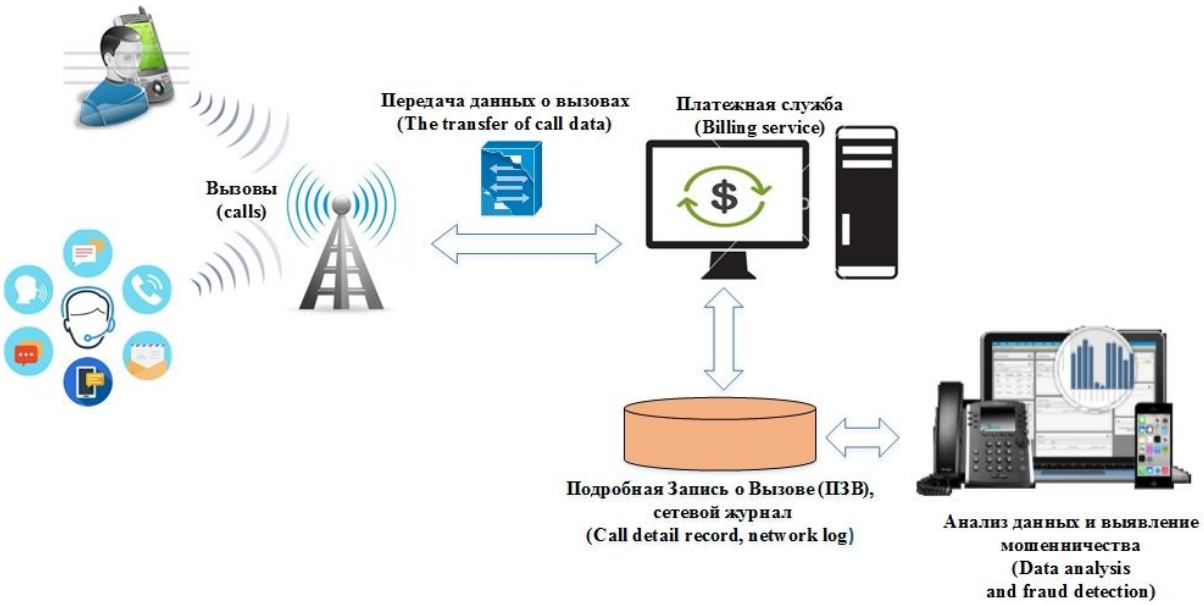
Чан Ван Фу, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 28, e-mail: vanphu.vstu.russia@gmail.com

Кравец Алла Григорьевна, доктор технических наук, профессор, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. Ленина, 28, e-mail:agk@gde.ru

В статье представлены результаты исследования в области обнаружения мошенничества потребителей при использовании ими услуг телекоммуникационных компаний. Основная идея заключается в превентивном (проактивном) обнаружении ситуаций, характеризующих использование потребителями мошеннических технологий. Предложен подход к проактивному обнаружению мошенничества, основанный на использовании ансамблей алгоритмов машинного обучения. В качестве выборки для обучения моделей предложено использовать синтетические данные в формате подробных записей о вызовах. Для выбора наилучшего алгоритма обнаружения мошеннических действий осуществлялся поиск гиперпараметров с оценкой эффективности на кроссвалидационной выборке. По результатам сравнения эффективности методов в качестве оптимального был выбран метод опорных векторов (SVM). Представлена архитектура распределенной системы обнаружения мошенничества потребителей услуг телекоммуникационных компаний, в которую интегрированы компоненты, реализующие предложенный метод.

Ключевые слова: телекоммуникационное мошенничество потребителей, обнаружение мошенничества, машинное обучение, обработка данных в реальном времени, пакетная обработка данных, методы интеллектуального анализа данных, методы принятия решений

Графическая аннотация (Graphical annotation)



[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 16-37-60066_мол_а_дк, № 15-07-06254 А и гранта Президента МД-6964.2016.9.

DEVELOPMENT OF A PROACTIVE METHOD FOR FRAUD DETECTIONIN IN TELECOMMUNICATION COMPANY

The article has been received by editorial board 14.11.2016, in the final version – 25.11.2016.

Nguyen Tuan Anh, post-graduate student, Volgograd State Technical University, 28 Lenin Av., Volgograd, 400005, Russian Federation, e-mail: anhtuank37@gmail.com

Shcherbakov Maksim V., D.Sc. (Engineering), Professor, Volgograd State Technical University, 28 Lenin Av., Volgograd, 400005, Russian Federation, e-mail:maxim.shcherbakov@gmail.com

Tran Van Phu, post-graduate student, Volgograd State Technical University, 28 Lenin Av., Volgograd, 400005, Russian Federation, e-mail: vanphu.vstu.russia@gmail.com

Kravets Alla G., D.Sc. (Engineering), Professor, Volgograd State Technical University, 28 Lenin Av., Volgograd, 400005, Russian Federation, e-mail: agk@gde.ru

The paper presents the results of research on consumer fraud detection in telecom. The basic idea is in proactive detection of fraud situations based on users' behaviour analysis. The approach based on the assembling of machine learning algorithms is proposed. Synthetic data sets containing Call Detail Records are used for training and testing algorithms. The choice of the best algorithm is made based on searching of hyperparameters with cross-validation. It is shown, that SVM is the most efficient algorithm for the certain data set. The distributed architecture of the fraud detection system was presented which implements the described approach.

Keywords: telecommunication fraud, fraud detection, machine learning, real time processing, batch processing, big data technique, data mining

Ведение. Интенсивное использование населением и организациями телекоммуникационных услуг (ТКУ) в развитых странах мира сопровождается ростом угроз для финансовой и информационной безопасности пользователей таких услуг [1–4] и телекоммуникационных компаний (ТКК). Хотя эти вопросы в существующих публикациях (например, в [6, 7, 9]) и рассматриваются, но некоторые направления остаются исследованными недостаточно полно. Поэтому целью настоящей статьи был анализ возможностей создания и использования программных средств, позволяющих выявлять факты мошенничества потребителей ТКУ с использованием проактивного подхода, основанного на машинном обучении.

Общая характеристика проблематики работы. Мошенничество потребителей все чаще становится сдерживающим фактором в отношении функционирования и развития предприятий, оказывающих ТКУ [18, 19, 22, 23]. В последние годы имеет место интенсивное развитие информационных технологий и увеличение количества интеллектуальных устройств, применяемых пользователями ТКУ. Одновременно наблюдается также рост потерь прибыли ТКК, связанный с фактами применения пользователями методов телекоммуникационного мошенничества (ТКМ). В телекоммуникационной отрасли мошенничество может быть определено как соответствующее наличию любого из следующих признаков:

- использование ТКУ без их оплаты пользователями (кража услуг связи);
- совершение других видов правонарушений с использованием услуг связи (в т.ч. кража со счетов других пользователей средств, находящихся на их балансах).

В первом случае потерпевшей стороной являются только ТКК (поставщики услуг связи), во втором – также и пользователи ТКУ.

По данным международной ассоциации по противодействию мошенничеству в сетях связи (CFCA) за 2013 г., общемировые потери операторов связи (ОС) от мошеннических действий составили $ 46,3 млрд, что на 15 % больше, чем в 2011 г. В среднем потери одного ОС от мошенников в 2013 г. составляли 2,09 % годового дохода, что больше, чем в 2011 г. Таким образом, наблюдается тенденция к увеличению потерь [10, 27]. Следует отметить, что пользователями часто используются и другие сервисы на основе ТКУ: финансовые, управленческие и т.д. В [30] была сделана попытка классификации видов ТКМ. При этом были выделены следующие основные классы мошеннических действий.

(1) Контрактное мошенничество (subscription fraud) – преднамеренное указание потребителем (абонентом) неверных данных при заключении контракта или невыполнение им контрактных условий оплаты. В этом случае нарушитель с самого начала не планирует платить за услуги или же в какой-то момент времени отказывается от их оплаты. Разновидностью этого вида ТКМ может быть ориентация потребителей лишь на частичную оплату ТКУ.

(2) Хакерское мошенничество (hacking fraud) – проникновение в компьютерную систему безопасности ТКК для удаления механизмов защиты или переконфигурирования этой системы с целью несанкционированного использования сети. Также мошенник (хакер) может использовать эту службу одновременно с абонентом, который оплачивает полученные им услуги. При этом хакером осуществляется дополнительное снятие средств в его пользу с баланса добросовестного плательщика.

(3) Техническое мошенничество (technical fraud) – неправомерное изготовление (клонирование) телефонных трубок или платежных телефонных карточек с фальшивыми идентификаторами абонентов,

номеров и платежных отметок. Специально отметим также мошеннические действия с терминальными системами [8], направленные на получение PIN-кодов пользователей, «задержку» в таких системах пластиковых карточек и пр.

Целесообразно отдельно выделить внутрикорпоративное техническое мошенничество. В этом случае мошенник пользуется ТКУ по сниженной цене – за счет незаконного доступа к корпоративной биллинговой системе. При соблюдении мошенником некоторых мер осторожности этот вариант мошенничества является наиболее сложным для обнаружения.

(4) Процедурное мошенничество (procedural fraud) – неправомочное вмешательство в бизнес-процессы (например, связанные с использованием биллинговых систем) с целью уменьшения оплаты услуг связи.

Специально отметим, что ТКМ может быть направлено и на несанкционированное получение информации о действиях клиентов ТКК, без нанесения им непосредственного финансового ущерба. Например, это может касаться следующего: получения персональной информации о клиентах, занесенной в базы данных ТКК; несанкционированного доступа к сведениям о номерах входящих и исходящих звонков клиентов ТКК, продолжительностях их переговоров и пр. При этом потенциально нарушение «информационной безопасности» клиентов может наносить вред не только им, но и ТКК (потеря репутации и, как следствие, клиентуры; необходимость возмещения ущербов клиентам по судебным искам и пр.).

Среди описанных четырех вариантов ТКМ наиболее распространенными являются контрактное и хакерское мошенничество. Последний тип ТКМ считается наиболее изученным.

Таким образом, проблема ТКМ является достаточно острой и требует повышения эффективности существующих (применяемых) мер. С учетом использования мошенниками различных способов (в т.ч. достаточно сложных) актуальной является задача обнаружения ТКМ с использованием методов интеллектуального анализа данных. При этом особое внимание целесообразно уделить проактивным средствам обнаружения ТКМ. К таким средствам относится, в частности, автоматизированная идентификация ситуаций, характеризующих возможные мошеннические действия пользователей ТКУ. В свою очередь, идентификация включает прогнозирование ситуаций мошенничества пользователя или группы пользователей ТКУ.

В настоящей работе рассмотрена идентификация одного из типов хакерского мошенничества – суперпозиционное мошенничество (СПМ – от англ. superimposed fraud). В этом случае мошенники используют учётную запись реального человека (пользователя услуг ТКК) для получения «за его счет» ТКУ [16, 30]. Такое мошенничество осуществляется, как правило, на уровне отдельных вызовов, а мошеннические вызовы в системах учета оказываются смешанными с легитимными. Полезность результатов, описанных в данной статье, заключается в обосновании целесообразности использования нового метода обнаружения ТКМ. Он отличается применением технологий машинного обучения – это позволяет минимизировать участие эксперта в формировании совокупности признаков выявления ТКМ. Кроме того, этот метод позволяет идентифицировать новые ситуации (не зафиксированные в прошлом), т.е. обладает определенными эвристическими возможностями.

Постановка задачи. Пусть имеются «Подробные Записи о Вызовах» (ПЗВ) абонентов (CallDetailRecords, CDR), характеризуемые следующим набором атрибутов: CDR = *<ID, CallNum, RcNum, StartTime, EndTime, Duration, Code, ChCos, Call_PR>*, где *ID* – идентификатор записи о конкретном вызове; *CallNum*- номер телефона вызывающего абонента; *RcNum* – номер телефона адресата (вызываемого абонента); *StartTime* – время начала разговора; *EndTime* – время окончания разговора; *Duration* – продолжительность соединения при разговоре (в минутах); *Code* – код страны, в которую абонент звонил; *ChCos* – цена звонка (рублей/минуту); *Call_PR* – стоимость звонка, заносимая в биллинговую систему ТКК.

Таким образом, каждая ПЗВ – это данные о событии, генерируемые в момент совершения пользователем вызова или использования другой ТКУ. Такими другими услугами могут в частности быть следующие: доступ пользователя в Интернет; проверка им «ящика» электронной почты; автоматизированное переключение абонента с одной базовой станции на другую в процессе его перемещения в пространстве. Приведенные выше регистрируемые параметры ПЗВ абонента заносятся в базу данных и в дальнейшем используются для анализа проведенных им (или с участием его устройства связи) операций. Пример фрагмента сгенерированных ПЗВ представлен в таблице 1.

Таблица 1 – **Пример сгенерированных ПЗВ (фрагмент базы данных)**

	id	Start_time	duration	area_code	charges_cost	call_price
0	655	5/30/2016 17:11	3.09	KZ	20.0	61.800
1	173	5/30/2016 17:16	0.61	UK	34.5	21.045
2	253	5/30/2016 17:18	0.74	VN	50.0	37.000
3	53	5/30/2016 17:21	2.03	RUS	12.0	24.060
4	213	5/30/2016 17:24	1.47	UK	34.5	50.715

Примечание: «id» – идентификатор абонента, «start_time» – время начала оказания услуги, «duration» – продолжительность (в мин.), «area_code» – код области, «charges_cost» – стоимости услуги, «call_price» – общая стоимость.

Далее будут рассмотрены две задачи обнаружения мошенничества.

• Задача З-1. Идентификация клиентов, которые относятся к возможным мошенникам (ВМ).

• Задача З-2. Прогнозирование поведения ВМ-клиентов, т.е. поддержка принятия проактивных решений для предотвращения возможных негативных последствий их действий для ТКК и ее клиентов.

На практике эти две задачи взаимосвязаны. Решение задачи З-1 помогает специалистам минимизировать время обнаружения пользователей-нарушителей на основе выявления их аномальных действий. Фактически эта задача может быть рассмотрена как задача идентификации аномалий в данных на основе задаваемых экспертами правил, характеризующих проявления ТКУ-мошенничества.

На основе выполнения задачи З-1 в базу данных ТКК заносится соответствующая информация. Эта информация может быть рассмотрена в дальнейшем как выборка данных. На основе указанной выборки может быть проведено обучение модели в рамках машинного обучения. Таким образом, правила выявления мошеннических действий могут быть получены автоматически – в результате использования алгоритмов машинного обучения.

Обзор литературных источников по теме статьи. В [6, 9, 18, 22, 23] представлены основные тенденции разработки систем обнаружения мошенничества, а также используемые в таких системах методы, основанные на применении перспективных технологий в области телекоммуникации. Особым образом в [6, 12, 13] выделены современные системы обнаружения ТКМ на основе обработки больших объемов данных, а также отмечены недостатки и сложности при создании таких систем.

Для создания систем успешного обнаружения ТКМ нужно решить следующие подзадачи:

• разработать методы сбора данных и формирования выборки данных с различными профилями поведения пользователей; методы обнаружения ТКМ, использующие модели машинного обучения;

• разработать методы поиска гиперпараметров моделей машинного обучения обнаружения ТКМ – для поиска наилучшей модели;

• реализовать эти методы в виде эффективной архитектуры системы, предназначенной для решения задач как в потоковом, так и в пакетном режимах обработки информации.

Рассмотрим методы и алгоритмы машинного обучения при решении задач идентификации ТКМ. В первую очередь следует определиться: имеется ли размеченная обучающая выборка или нет. Размеченная выборка характеризуется тем, что объекты уже классифицированы как «мошеннические» и «не мошеннические».

В случае отсутствия размеченной обучающей выборки используются подходы для идентификации ТКМ на основе методов кластеризации [15]. К ним относят, в частности, метод *k*-средних, метод *k*-ближайших соседей (k-nearest neighbors algorithm, k-NN). В [14] представлен подход, предназначенный для обнаружения мошенничества с использованием алгоритмов генетической кластеризации – с целью решения задачи оценки и обеспечения качества исходных данных.

При наличии размеченной выборки применяются линейные методы и нелинейные модели машинного обучения. В [19, 30] использован наивный байесовский классификатор. Метод опорных векторов (Support Vector Machines, SVM) также относится к одним из популярных методов. В [26] показана возможность применения SVM для обнаружения мошенничества с кредитными картами и в других областях. В [7, 26] также показана высокая точность обнаружения мошенничества при использовании методов SVM – в том числе SVM с радиально базисной функции (РБФ) ядра, SVM с полиномиальными ядрами, SVM с сигмоидным ядром, SVM с линейным ядром. В [29] авторами созданы новые подходы на основе объединения различных методов SVM и технологии нечёткой кластеризации.

Кроме перечисленных алгоритмов, в практике часто используются методы машинного обучения на основе деревьев решений [8, 25, 26]. Это могут быть деревья классификаций, регрессионные деревья и другие. В ряде работ предложены различные архитектуры для программных систем обнаружения мошенничества, а также методы обработки данных [5, 23].

Таким образом, на основе преимуществ перечисленных методов имеется возможность применения их в комплексной задаче обнаружения мошенничества. В настоящей статье предлагается использовать подход, основанный на формировании ансамблей моделей [5]. Такой подход позволяет эффективно использовать несколько разных моделей для достижения лучшей точности прогнозирования, по сравнению с применением их по отдельности. Основная цель при проектировании ансамблей моделей такая же, как при привлечении для принятия решений группы (комитета) людей: каждый член комитета должен быть настолько компетентным, насколько это возможно; при этом участники группы должны функционально дополнять друг друга. Если участники не являются взаимодополняющими, т. е. если они всегда согласны друг с другом, тогда комитет не нужен и достаточно одного любого члена. Если участники дополняют друг друга, то в случаях, когда один или несколько членов допускают ошибки, велика вероятность того, что остальные участники могут эту ошибку исправить.

Метод обнаружения мошенничества. Профили пользователей создаются в результате анализа их действий, содержат краткие описания их поведения с учётом времени и местонахождения.

Первый профиль (из числа рассматриваемых далее) создан на основе анализа данных о поведении пользователей, накапливаемых еженедельно. Профиль состоит из семи полей.

Профиль1 = <*'calls_mean','calls_std','calls_max','dur_mean','dur_mean_max', 'dur_max', 'dur_std','cost_max','int_do_rate'*>,

где: '*calls_mean'* – среднее арифметическое количество вызовов в неделю;
'calls_std' – стандартное отклонение количества вызовов в неделю;
'calls_max' – максимальное количество вызовов;
'dur_mean' –средняя арифметическая продолжительность вызовов в неделю;
'dur_mean_max' – максимальное значение из средних арифметических продолжительностей вызовов, которые выполняются по будним дням;
'dur_max' – максимальная продолжительность одного вызова в неделю;
'dur_std' – стандартное отклонение для продолжительности вызовов в неделю;
'cost_max' – максимальная стоимость одного вызова в неделю;
'int_do_rate' – процент международных звонков от общего числа звонков за неделю.

Второй профиль – это описание детального повседневного поведения пользователей, которое создаётся путём агрегации количества вызовов в день, а также их продолжительностей по вызываемым направлениям:

Профиль2 =
<*'w_cost_max' 'w_calls_count' 'w_dur_max' 'w_dur_mean' 'w_dur_std'*
'af_cost_max' 'af_calls_count' 'af_dur_max' 'af_dur_mean' 'af_dur_std'
'ni_cost_max' 'ni_calls_count' 'ni_dur_max' 'ni_dur_mean' 'ni_dur_std'
'int_do_rate'>,

где '*w_cost_max*– максимальная стоимость одного вызова по рабочим часам (с 7:00 до 17:00) (w);
'w_calls_count' – количество вызовов по рабочим часами (w);
'w_dur_max' – максимальная продолжительность одного вызова по рабочим часами;
'w_dur_mean' – средняя арифметическая продолжительность по рабочим часами;
'w_dur_std' – стандартное отклонение для продолжительностей вызовов по рабочим часами;
'af_cost_max' - максимальная стоимость одного вызова в вечерние часы (с 17:00 по 23:00);
'af_calls_count' – количество вызовов в дневные часы;
'af_dur_max' – максимальная продолжительность одного вызова в дневные часы;
'af_dur_mean' – средняя арифметическая продолжительность вызовов в дневные часы;
'af_dur_std' – стандартное отклонение для продолжительностей вызовов в дневные часы;
'ni_cost_max' – максимальная стоимость одного вызова в ночные часы (с 23:00 до 07:00);
'ni_calls_count' – количество вызовов в ночные часы;
'ni_dur_max' – максимальная продолжительность одного вызова в ночные часы;
'ni_dur_mean' – средняя арифметическая продолжительность вызовов в ночные часы;
'ni_dur_std' – стандартное отклонение для продолжительности вызовов в ночные часы;
'int_do_rate' – процент международных звонков от общего количества звонков за сутки.

На рисунке 2 представлена визуализация искусственно сгенерированных данных о звонках пользователей по дням недели.

Отметим, что поведение мошенников отличается частотой и средней продолжительностью звонков (рис. 2б) от поведения «нормальных» пользователей (рис. 2а). Такая ситуация характеризует тип СПМ, при котором неправомерные действия «наложены» на действия обычных клиентов (пользователей).

Рассмотрим предлагаемый метод, включающий в себя два основных этапа: построение модели и ее применение.

На этапе построения модели выполняются следующие действия.

1. Сбор данных и их структурирование в формате, описанном выше.
2. Предобработка исходных данных с целью обнаружения пропущенных значений для атрибутов ‘duration’, ‘call_price’и замены их средними значениями.
3. Формирование выборок данных для обучения модели.

a. Формирование первой выборки данных (маркировка profile1_ds), включающей данные в соответствии с «профилем 1», которые размечены для каждого пользователя: 0 – обычный пользователь, 1 – мошенник.

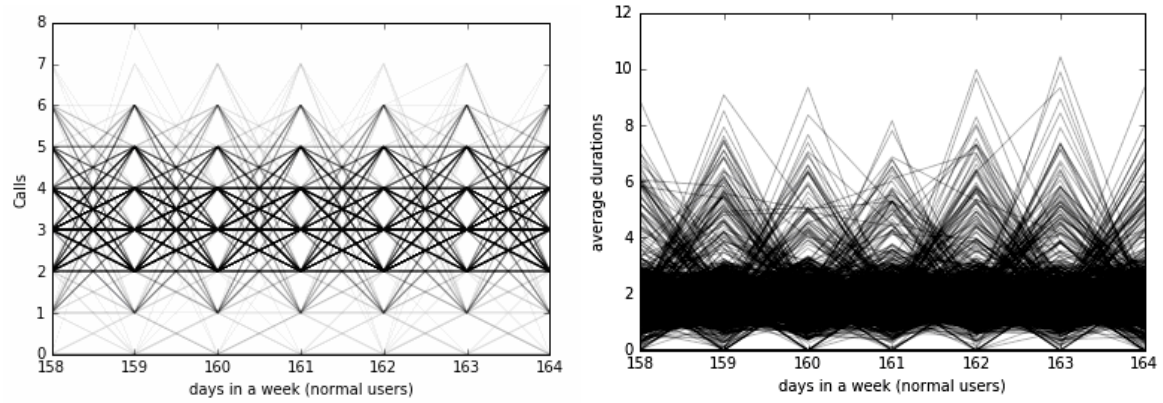
b. Формирование второй выборки данных (маркировка profile2_ds), включающей данные в соответствии с «профилем 2» и размеченные для каждого пользователя: 0 – обычный пользователь, 1 – мошенник.

4. Формирование обучающей, кроссвалидационной и тестовой выборок. В этом случае осуществляется разбиение выборки в соответствии с пропорцией: 80 % для обучения, а 20 % оставшихся данных – для тестирования.

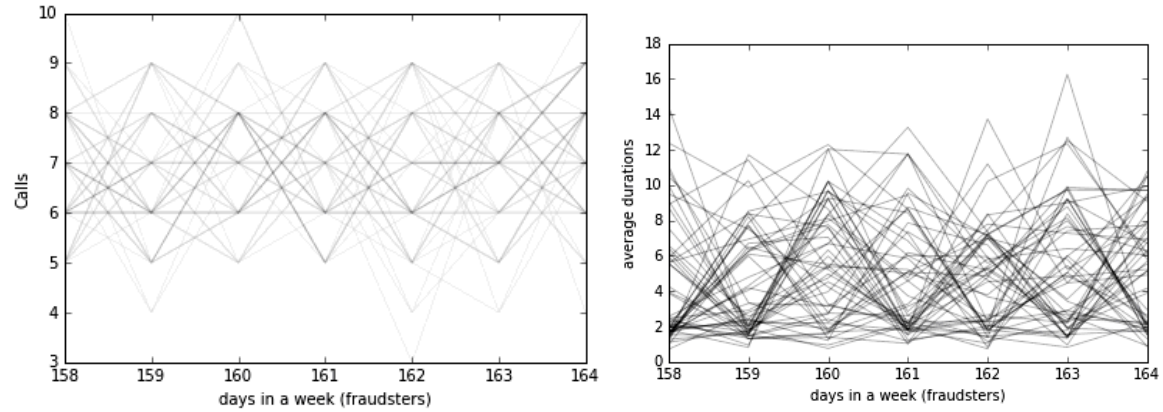
5. Настройка и оценка моделей. Настройка моделей осуществляется на выделенной обучающей выборке исходя из предопределенного числа моделей. Для поиска наилучшей модели используется приём поиска гиперпараметров по сетке (GridSearch) с проверкой на кроссвалидационных выборках с установленным числом разбиения (по умолчанию – на 10). Настройка осуществляется отдельно для каждого из наборов данных.

6. Выбор наилучших моделей (с наилучшими значениями гиперпараметров) на основе значения точности для выборок profile1_ds и profile2_ds. Затем производится дообучение этих моделей на всём (100 %) наборе данных.

7. Сохранение параметров лучшей модели для последующего обнаружения ТКМ.



(а)



(б)

Рис. 2 – Визуализация частоты звонков и средней продолжительности разговоров для «нормальных пользователей» (а) и «мошенников» (б) по «недельным данным». Обозначения на рисунке: Calls – число звонков в течение суток, Average Duration – средняя продолжительность звонков (мин.). По оси абсцисс – номера дней (суток), отсчитанные с начала года

Этап применения модели:

1. Загрузка параметров модели для соответствующей задачи обнаружения ТКМ.
2. Предобработка исходных данных в режиме реального времени.
3. Подготовка входных векторов для представления их в качестве входов в моделях.
4. Расчет выходных значений для полученных моделей.
5. Генерация события-оповещения, если спрогнозированное поведение пользователя относится к МП.

Архитектура системы обнаружения ТКМ. В настоящее время существуют различные подходы к разработке архитектуры системы поддержки принятия решений на основе обработки больших данных. Известен подход на основе лямбда-архитектуры, комбинирующий в себе обработку данных в потоковом режиме (режиме реального времени) и пакетную обработку [21]. В рамках такой архитектуры для реше-

ния задачи сбора и передачи данных часто используются решения на основе Flume, Kafka, Spark, SparkStreaming [24,28]. Эти технологии позволяют реализовывать сбор и передачу данных в режиме реального времени, а также их потоковую обработку. Для решения задач хранения данных и их обработки в пакетном режиме применяется технологический стек Apache Hadoop с распределенной файловой системой (Hadoop Distributed File System), предназначенной для реализации программ в концепции MapReduce [17]. Кроме того, могут быть использованы СУБД CASSANDRA, HBASE [11]. В частности, в [24] представлены решения рассматриваемых в статье задач в режиме реального времени с использованием фреймворка SparkStreaming.

Архитектура предлагаемой авторами настоящей статьи программной системы обнаружения ТКМ состоит из пяти подсистем:

- S1 – подсистема источников данных;
- S2 – подсистема сбора данных;
- S3 – подсистема обработки и анализа данных;
- S4 – подсистема хранения результатов обработки данных;
- S5 – подсистема визуализации результатов.

Исходные данные представляются собой подробные записи о вызовах клиентов. На практике такие данные сохраняются во внутренней базе данных ТКК, то есть они не находятся в общем доступе [13]. Поэтому, для проведения испытаний была разработана подсистема генерации ПЗВ-логов клиентов на основе технологии Kafka [20]. Kafka-кластер состоит из множества брокеров для сбора данных, потребителей (Customers) и производителей (Producers). Потребители формируют подписку на определенные брокеры для сбора данных. В подсистеме S1 было создано несколько производителей для генерации данных, а в подсистеме S2 – несколько потребителей данных. Описанная модель позволяет эмулировать работу по сбору данных в режиме реального времени.

В подсистеме S3 процессы обработки данных выполняются в двух режимах: (1) пакетном и (2) потоковом.

В пакетном режиме осуществляется построение модели обнаружения ТКМ на основе технологического стека Hadoop с использованием методов машинного обучения [18]. Данные загружаются из Hadoop HDFS для обработки. Полученная модель сохраняется в подсистеме S4.

В потоковом режиме осуществляется классификация событий на основе полученной в пакетном режиме модели. Также происходит обновление параметров модели на основе новых входных данных. В подсистеме S5 выполняется визуализация результатов в формате отчета «Прогнозирование потенциального мошенничества».

Результаты испытания метода и их обсуждение. Для решения задачи обнаружения мошенничества в этой работе использованы 5 моделей и алгоритмов машинного обучения: алгоритм логистической регрессии (LR); деревья классификации и регрессии (CART); SVM; наивный байесовский классификатор (NB), метод *k*-ближайших соседей (*k*-NN). В ходе испытания системы были сгенерированы выборки данных, содержащие по 146240 записей, структура которых соответствует схемам «Профиль1» и «Профиль2», описанных выше.

После обучения моделей была проведена оценка их достоверности (отношение суммы истинно-положительной и истинно-отрицательной оценок к сумме истинно-положительной, истинно-отрицательной и ложно-отрицательной оценками). Результаты работы для тестовой выборки представлены в таблице 2.

Таблица 2 – **Результаты работы алгоритмов на обучающей выборке**

Обозначения алгоритмов	Средняя достоверность
LR	0.926
k-NN	**0.926**
CART	0.841
NB	0.919
SVM	**0.927**

Для определения наилучшего алгоритма был выполнен поиск лучшей комбинации гиперпараметров. Для алгоритма *k*-NN гиперпараметр '*n_neighbors*' – число ближайших соседей, по умолчанию равен 7. Авторами были получены результаты для всех нечётных значений этого параметра от 1 до 21 на кроссвалидационной выборке данных. Параметр разбиения кроссвалидационной выборки данных *k_fold* = 10 характеризует число блоков данных, на которые разбивается выборка. При этом последовательно модель обучается на *k-i* блоке, а ее качество проверяется на *i*-м блоке. Для *k*-NN алгоритма лучшие результаты классификации составили 0.926 со значением параметра *n_neighbors* = 5.

Также был проведен поиск лучших значений двух основных параметров алгоритма SVM: (1) значение коэффициента *C*, который позволяет регулировать отношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки; (2) тип ядра. По умолчанию для SVM ис-

пользуются РБФ-ядра со значением C = 1.0. Как и для предыдущего алгоритма был выполнен поиск на кроссвалидационной выборке данных с таким же параметром разбиения. Были использованы несколько простых функций ядра, значения C устанавливалось меньше и больше 1.0.

Алгоритм SVM со значением C = 0.9 и сигмоидальным видом ядра показал наилучший результат с точки зрения точности классификации. Точность алгоритма SVM была выше точности алгоритма *k*-NN, хотя и незначительно. Поэтому алгоритм SVM был выбран как основной для генерации модели обнаружения ТКМ. Средние значения мер точности по этому алгоритму для тестовых выборок оказались следующие: точность (precision) = 0.78, полнота (recall) = 0.83.

Выводы. В данной работе рассмотрена проблема обнаружения мошенничества при использовании ТКУ. Предложен метод обнаружения ТКМ, отличающийся применением нескольких алгоритмов машинного обучения и выбором наилучшего из них. Это позволяет минимизировать участие эксперта в формировании условий обнаружения МП пользователей и идентифицировать новые ситуации, не зафиксированные в прошлом.

Целесообразно использовать несколько алгоритмов машинного обучения, рассматриваемых в данной статье, – для выбора наилучшего в соответствии с различными критериями качества. Кроме того, поиск на сетке лучшей комбинации гиперпараметров позволяет исключить возможность пропуска лучшей модели. В частности, в результате проведенных экспериментов был выбран алгоритм SMV с сигмоидальным ядром, позволивший получить точность 0.78 на тестовой выборке данных по «профилю 2».

Представленная архитектура системы обнаружения ТКМ позволяет реализовывать алгоритмы для использования их как в пакетном (этап построения моделей обнаружения ТМК), так и в потоковом (этап применения моделей обнаружения ТМК) режимах.

Список литературы

1. Брумштейн Ю. М. Влияние информационно-телекоммуникационной компетентности граждан на направления и эффективность использования ими услуг сотовой связи / Ю. М. Брумштейн, А. Б. Кузьмина // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 3. – С. 48.

2. Буй Нгок Зыонг. Безопасная аутентификация в системе управления корпоративной мобильностью / Нгок Зыонг Буй, А. Г. Кравец, Ле Тхань Тунг Нгуен // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. – 2015. – № 13 (177). – С. 45–51.

3. Нгуен Туан Ань. Архитектура обнаружения мошенничества в телекоммуникационном предприятии с Hadoop / Туан Ань Нгуен, В. А. Камаев, М. В. Щербаков // Мир науки и инноваций. – 2015. – Вып. 2, т. 2 «Технические науки». – С. 75–78.

4. Учаев Д. Ю. Анализ и управление рисками, связанными с информационным обеспечением человеко-машинных АСУ технологическими процессами в реальном времени / Д. Ю. Учаев, Ю. М. Брумштейн, И. М. Ажмухадедов, О. М. Князева, И. А. Дюдиков // Прикаспийский журнал: управление и высокие технологии. – 2016. – № 2. – С. 161–175.

5. Чан Ван Фу. Обзор архитектур систем поддержки принятия решений, использующих аналитику данных в режиме реального времени / Ван Фу Чан, М. В. Щербаков, Туан Ань Нгуен // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. – 2016. – № 3 (182). – С. 95–100.

6. Abdallah A. Fraud detection system: A survey / A. Abdallah, M. A. Maarof, A. Zainal // Journal of Network and Computer Applications. – 2016 – Vol. 68. – P. 90–113.

7. Abdelhamid D. Automatic Bank Fraud Detection Using Support Vector Machines / D. Abdelhamid, S. Khaoula, O. Atika // The International Conference on Computing Technology and Information Management (ICCTIM2014). – 2014. – P. 10–17.

8. Anis M. A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection / M. Anis, M. Ali, A. Yadav // International Journal of Economics, Commerce and Management. – 2015. – Vol. III, № 12.

9. Bolton. Statistical fraud detection: a review / Bolton, J. Richard, Hand, J. David // Stat. Sci. – 2002. – № 17 (3). – P. 235–255.

10. Communications Fraud Control Association. – Режим доступа: http://www.cfca.org (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

11. Celko J. Complete Guide to NoSQL_ What Every SQL Professional Needs to Know about Non-Relational Databases / J. Celko // Morgan Kaufmann. – 2013. – 244 p.

12. Chen F. Data Mining for the Internet of Things Literature Review and Challenges / F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, X. Rong // Int. J. Distrib. Sens. Netw. – 2015. – Vol. 15, № 12. – P. 1–14.

13. Cox K. C. Visual data mining: Recognizing telephone calling fraud / K. C. Cox, S. G. Erick, G. J. Wills // Data Mining and Knowledge Discovery. – 1997. – № 1. – P. 225–231.

14. Ganji V. R. Credit card fraud detection using anti *k*-nearest algorithm / V. R. Ganji, S. N. P. Mannem // IJCSE. – 2012. – Vol. 4, № 6.

15. Goldstein M. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data / M. Goldstein, S. Uchida. – Режим доступа: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836738/pdf/pone.0152173.pdf (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

16. Gosset P. Classification, detection and prosecution of fraud in mobile networks/ P. Gosset и M. Hyland // Proceedings of ACTS Mobile Summit. – Sorrento, Italy, June 1999.

17. HadoopApache. – Режим доступа: http://hadoop.apache.org (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

18. Hoath P. Telecoms Fraud, The Gory Details / P. Hoath // Computer Fraud & Security. – 1998. – Vol. 98, № 1. – P. 10–14.

19. Kabari L. G. Telecommunications Subscription Fraud Detection Using Naïve Bayesian Network / L. G. Kabari, D. N. Nanwin, E. U. Nquoh // IIARD Int. J. Com. Sci. Sta. – 2016. – Vol. 2, № 2.

20. KafkaApache. – Режим доступа: https://kafka.apache.org (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

21. Lambda Architecture Repository. – Режим доступа: http://lambda-architecture.net (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

22. Laleh N. A Taxonomy of Frauds and Fraud Detection Techniques / N. Laleh, A. M. Azgomi // ICISTM. – 2009. – Vol. 31. – P. 256–267.

23. Moudani W. Fraud Detection in Mobile Telecommunication / W. Moudani, F. Chakik // LNSE. – 2013. – Vol. 1. – P. 75–79.

24. Nabi Z. Pro Spark Streaming- The Zen of Real-time Analytics using Apache Spark / Z. Nabi // Apress. – 2016. – 252 p.

25. Rokach L. Data mining with decision trees – Theory and Applications 2nd Edition / L. Rokach, O. Maimon // Series in Machine Perception and Artificial Intelligence. – 2015. – Vol. 81. – 328 p.

26. Sahin Y. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines / Y. Sahin, E. Duman // Pro. Int. Mul. Eng. Com. Scien. IMECS. – 2011. – Vol. 1.

27. Singh P. Fraud Detection by Monitoring Customer Behavior and Activities / P. Singh, M. Singh // International Journal of Computer Applications. – 2015. – Vol. 111, № 11.

28. SparkApache. – Режим доступа: http://spark.apache.org (дата обращения 17.10.2016), свободный. – Заглавие с экрана. – Яз. англ.

29. Subudhi S. Use of fuzzy clustering and support vector machine for detecting fraud in mobile telecommunication networks / S. Subudhi, S. Panigrahi // International Journal of Security and Networks. – 2016. – Vol. 11, № 1/2. – P. 3–11.

30. Yusoff. Fraud detection in telecommunication industry using Gaussian mixed model / Yusoff, I. M. Mohd, Mohamed, Ibrahim, Bakar, Mohd. A. Rizam // Proceedings of the 2013 International Conference on Research and Innovation Information Systems. – P. 27–32.

References

1. Brumshteyn Yu. M., Kuzmina A. B. Vliyanie informatsionno-telekommunikatsionnoy kompetentnosti grazhdan na napravleniya i effektivnost ispolzovaniya imi uslug sotovoy svyazi [Influence of information and telecommunication citizens competence at the directions and efficiency of cellular communication services usage]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Control and High Technologies], 2014, no. 3, pp. 48–67.

2 Buy Ngok Zyong, Kravets A. G., Le Tkhan Tung Nguyen. Bezopasnaya autentifikatsiya v sisteme upravleniya korporativ-noy mobilnostyu [Safe authentication in management of system corporate mobility]. *Izvestiya VolgGTU. Ser. Aktualnye problemy upravleniya, vychislitelnoy tekhniki i informatiki v tekhnicheskikh sistemakh* [Proceedings of the VolgGTU, Series "Actual Problems of Management, Computing Hardware and Informatics in Engineering Systems"], 2015, no. 13 (177), pp. 45–51.

3. Nguyen Tuan An, Kamaev V. A., Shcherbakov M. V. Arkhitektura obnaruzheniya moshennichestva v telekommunikatsionnom predpriyatii s Hadoop [Architecture of fraud detection in the telecommunication company with Hadoop]. *Mir nauki i innovatsiy* [The World of Science and Innovations], 2015, issue 2, vol. 2 «Technical sciences», pp. 75–78.

4. Uchaev D. Yu., Brumshteyn Yu. M., Azhmukhadedov I. M., Knyazeva O. M., Dyudikov I. A. Analiz i upravlenie riskami, svyazannymi s informatsionnym obespecheniem cheloveko-mashinnykh ASU tekhnologicheskimi protsessami v realnom vremeni [The risk-analysis and risk-management, for procedures connected with information support of man-machine automated systems, working in real time]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Control and High Technologies], 2016, no. 2, pp. 161–175.

5. Chan Van Fu, Shcherbakov M. V., Tuan An Nguyen. Obzor arkhitektur sistem podderzhki prinyatiya resheniy, ispol-zuyushchikh analitiku dannykh v rezhime realnogo vremeni [The overview of architecture of the systems for decision support, using data analytics in real time] . *Izvestiya VolgGTU. Ser. Aktualnye problemy upravleniya, vychislitelnoy tekhniki i informatiki v tekhnicheskikh sistemakh* [Proceedings of the VolgGTU, Series "Actual Problems of Management, Computing Hardware and Informatics in Engineering Systems"], 2016, no. 3 (182), pp. 95–100.

6. Abdallah A., Maarof M. A., Zainal A. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 2016, vol. 68, pp. 90–113.

7. Abdelhamid D., Khaoula S., Atika O. Automatic Bank Fraud Detection Using Support Vector Machines. *The International Conference on Computing Technology and Information Management (ICCTIM2014)*, 2014, pp. 10–17.

8. Anis M., Ali M., Yadav A. A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection. *International Journal of Economics, Commerce and Management*, 2015, vol. III, no. 12.

9. Bolton, Richard J., Hand, David J. Statistical fraud detection: a review. *Stat. Sci.*, 2002, no. 17 (3), pp. 235–255.

10. *Communications Fraud Control Association*. Available at: http://www.cfca.org (accessed 17.10.2016).

11. Celko J. Complete Guide to NoSQL_ What Every SQL Professional Needs to Know about Non-Relational Databases. *Morgan Kaufmann*, 2013. 244 c.

12. Chen F., Deng P., Wan J., Zhang D., Vasilakos A. V., Rong X. Data Mining for the Internet of Things Literature Review and Challenges. *Int. J. Distrib. Sens. Netw.*, 2015, vol. 15, no. 12, pp. 1–14.

13. Cox K. C., Erick S. G., Wills G. J. Visual data mining: Recognizing telephone calling fraud. *Mining and Knowledge Discovery*, 1997, no. 1, pp. 225–231.

14. Ganji V. R., Mannem S. N. P. Credit card fraud detection using anti *k*-nearest algorithm. *IJCSE*, 2012, vol. 4, no. 6.

15. Goldstein M., Uchida S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836738/pdf/pone.0152173.pdf (accessed 17.10.2016).

16. Gosset P., Hyland M. Classification, detection and prosecution of fraud in mobile networks. *Proceedings of ACTS Mobile Summit*, Sorrento, Italy, June 1999.

17. HadoopApache. Available at: http://hadoop.apache.org (accessed 17.10.2016).

18. Hoath P. Telecoms Fraud, The Gory Details. *Computer Fraud & Security*, 1998, vol. 98, no. 1, pp. 10–14.

19. Kabari L. G., Nanwin D. N., Nquoh E. U. Telecommunications Subscription Fraud Detection Using Naïve Bayesian Network. *IIARD Int. J. Com. Sci. Sta*, 2016, vol. 2, no. 2.

20. KafkaApache. Available at: https://kafka.apache.org (accessed 17.10.2016).

21. Lambda Architecture Repository. Available at: http://lambda-architecture.net (accessed 17.10.2016).

22. Laleh N., Azgomi A. M. A Taxonomy of Frauds and Fraud Detection Techniques. *ICISTM*, 2009, vol. 31, pp. 256–267.

23. Moudani W., Chakik F. Fraud Detection in Mobile Telecommunication. *LNSE*, 2013, vol. 1, pp. 75–79.

24. Nabi Z. Pro Spark Streaming- The Zen of Real-time Analytics using Apache Spark. *Apress*, 2016. 252 p.

25. Rokach L., Maimon O. Data mining with decision trees – Theory and Applications 2nd Edition. *Series in Machine Perception and Artificial Intelligence*, 2015, vol. 81. 328 c.

26. Sahin Y., Duman E. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. *Pro. Int. Mul. Eng. Com. Scien. IMECS,* 2011, vol. 1.

27. Singh P. Fraud Detection by Monitoring Customer Behavior and Activities / P. Singh, M. Singh // International Journal of Computer Applications – 2015 – Vol. 111 – № 11.

28. SparkApache. Available at: http://spark.apache.org (accessed 17.10.2016).

29. Subudhi S., Panigrahi S. Use of fuzzy clustering and support vector machine for detecting fraud in mobile telecommunication networks. *International Journal of Security and Networks*, 2016, vol. 11, no. ½, pp. 3–11.

30. Yusoff, Mohd I. M., Mohamed, Ibrahim, Bakar, Mohd. A. Rizam. Fraud detection in telecommunication industry using Gaussian mixed model. *Proceedings of the 2013 International Conference on Research and Innovation Information Systems*, pp. 27–32.

УДК [004.738.5+004.5]:[352+353]

ИНТЕРНЕТ-САЙТЫ ОРГАНОВ ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ В ПРИКАСПИЙСКИХ РЕГИОНАХ РОССИИ: АНАЛИЗ НОМЕНКЛАТУРЫ, ФУНКЦИОНАЛЬНОСТИ И ВОСТРЕБОВАННОСТИ РЕСУРСОВ

Статья поступила в редакцию 17.09.2016, в окончательном варианте – 28.11.2016.

Брумштейн Юрий Моисеевич, кандидат технических наук, доцент, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: brum2003@mail.ru

Васьковский Евгений Юрьевич, аспирант, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, e-mail: vaskovskiy_evgeniy@mail.ru

Горбачева Анна Николаевна, программист, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а, gorbacheva.ann@gmail.com

Рассмотрены особенности административно-территориального деления и соответствующих им органов исполнительной власти (ОИВ) России. Для регионов, прилежащих к Каспийскому морю, приведены численности населения и площади территорий. Показана роль использования интернет-сайтов в деятельности ОИВ этих регионов. Обоснована целесообразность анализа информационного присутствия в Интернете ОИВ регионов на трех иерархических уровнях: региональном (область, республика); районном; местном (муниципалитет, сельсовет и т.д.). Проанализированы номенклатура и особенности построения сайтов ОИВ в прикаспийских регионах России на разных иерархических уровнях, функциональность этих сайтов, подходы к их информационному наполнению, принципы организации интерфейсов с пользователями и др. Приведены таблицы с основными характеристиками сайтов рассматриваемых в статье ОИВ: времена открытия стартовых страниц; оценки количеств страниц на сайтах; количества внутренних ссылок (как показатели связанности размещенных на сайтах информационных материалов); количеств входящих и исходящих ссылок – как меры информационной связности сайтов с интернет-пространством; показатели посещаемости сайтов и др. Обоснована целесообразность нормирования некоторых из приведенных показателей для целей сравнения различных сайтов. Оценены частоты обновления информации на сайтах ОИВ различных иерархических уровней. Сделаны выводы о достоинствах и недостатках рассмотренных сайтов.

Ключевые слова: Россия, прикаспийские регионы, органы исполнительной власти, интернет-сайты, видимость сайтов, востребованность сайтов, управление востребованностью, информационные технологии, информационное присутствие, методы анализа, методы управления, поисковые системы